DOCUMENT RESUME

ED 217 063 TM 820 278

AUTHOR Fienberg, Stephen E.; Larntz, Kinley
TITLE Design and Analysis Problems Associated with Qualitative Data in Educational Research. Final Report.
INSTITUTION Minnesota Univ., St. Paul.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
PUB DATE 25 Jan 79
GRANT NIE-G-76-0094
NOTE 31p.

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Data Analysis; *Educational Research; Mathematical Models; *Research Problems; *Statistical Analysis
IDENTIFIERS Categorical Data; *Log Linear Models; *Qualitative Data

ABSTRACT

This research project addresses a series of methodological and theoretical statistical problems in the analysis of categorical data using loglinear and logistic response models, which grow directly out of problems in the study of the American educational system, and in basic educational research. The project focuses on the adaptation and development of general results on loglinear models to handle a wide variety of data-analytic situations involving qualitative data. Problems that are considered include: (1) modelling age, period, and cohort effects in the analysis of discrete archival data; (2) analysis and design problems related to the use of longitudinal and cross-sectional qualitative data; (3) simultaneous systems of logit models: (4) small sample properties of chi-square statistics; (5) models for mixtures of qualitative and quantitative variables; (6) asymptotic properties of chi-square statistics in large sparse contingency tables; (7) selection procedures for hierarchical loglinear model fitting; (8) the application of general loglinear model results to new problems, such as paired comparison models allowing for ties and order effects; (9) analysis of qualitative data from complex sample surveys, with special attention to cluster sampling; and (10) graphical methods for the presentation and analysis of multivariate data. (Author/GK)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED217063

FINAL REPORT

GRANT NO. NIE G-76-0094

DESIGN AND ANALYSIS PROBLEMS ASSOCIATED WITH QUALITATIVE DATA IN EDUCATIONAL RESEARCH

STEPHEN E. FIENBERG & KINLEY LARNTZ
UNIVERSITY OF MINNESOTA
ST. PAUL, MINNESOTA 55108

JANUARY 25, 1979

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

The research reported herein was performed pursuant to a grant with the National Institute of Education, U.S. Department of Health, Education and Welfare. Contractors undertaking such projects under Government sponsorship are encouraged to express freely their professional judgement in the conduct of the project. Points of view or opinions stated do not, therefore, necessarily represent official National Institute of Education position or policy.

TM 820 278

U.S. DEPARTMENT OF

HEALTH, EDUCATION AND WELFARE

NATIONAL INSTITUTE OF EDUCATION

BASIC SKILLS

ABSTRACT

This research project has addressed a series of methodological and theoretical statistical problems in the analysis of categorical data using loglinear and logistic response models, which grow directly out of problems in the study of the American educational system, and in basic educational research. The project has focused on the adaptation and development of general results on loglinear models to handle a wide variety of data-analytic situations involving qualitative data. Problems that have been considered included:

1. modelling age, period, and cohort effects in the analysis of discrete archival data,
2. analysis and design problems related to the use of longitudinal and cross-sectional qualitative data,
3. simultaneous systems of logit models,
4. small sample properties of chi-square statistics,
5. models for mixtures of qualitative and quantitative variables,
6. asymptotic properties of chi-square statistics in large sparse contingency tables,
7. selection procedures for hierarchical loglinear model fitting,
8. the application of general loglinear model results to new problems, such as paired comparison models allowing for ties and order effects,
9. analysis of qualitative data from complex sample surveys, with special attention to cluster sampling,
10. graphical methods for the presentation and analysis of multivariate data.

I. INTRODUCTION

(a) Background.

The analysis of cross-classified categorical data has occupied a prominent place in the statistical literature of the present century, but with a few exceptions the only techniques used in the analysis of frequency data arising from educational research have been those associated with the analysis of two-dimensional contingency tables and the calculation of chi-square statistics or measures of association. During the past 15 years advances in statistical theory and the ready availability of high-speed computers have led to major advances in the analysis of multidimensional categorical data via the use of loglinear and logit models. Recent books by Bishop, Fienberg, and Holland [1975], Cox [1970], Fienberg [1977], Gokhale and Kullback [1978], Haberman [1974], Lindsey [1973], and Plackett [1974] present detailed expositions of these new loglinear and logit model techniques, and these books in turn have stimulated further methodological research, and, more importantly, the application of the basic loglinear and logit model techniques.

Despite the lavish attention given to the analysis of cross-classified categorical data by statisticians and other methodological researchers, little effort has gone toward the design of experiments involving categorical response variates, and only slightly more toward the design of suitable observational studies (e.g. see the review by McKinlay [1975].

While general theoretical results on the use of loglinear and logistic response models for multidimensional categorical data are

available (e.g. see Haberman [1974]), the adaption of these general results to specific problems of widespread interest in the educational research community requires considerable effort. Moreover, many analytical statistical problems remain which require new theoretical results. Some of these problems are briefly described in Fienberg [1977].

(b) Results of Research Project.

This research project has provided the solutions to a series of methodological and theoretical statistical problems in the analysis of categorical data that grow directly out of problems in the study of the American educational system, and in basic educational research. The adaptation of the general loglinear model for widespread use by the educational research community requires considerable effort and has also generated many new problem areas. We have dealt with several critical aspects of this adaptation. Typical of this effort has been the preparation of a survey paper on the structure of the general loglinear model approach and its application to different problems, and the preparation of a brief paper on fitting and interpreting parameters in loglinear models.

Considerable interest has been focused on national trends in education, but since data are generally available only on a cross-sectional basis few authors have been able to simultaneously determine age, period, and cohort effects on educational attainment, especially when the latter is considered as a categorical response variable. We developed logistic response models, suitable for such problems, first for the dichotomous response case, and then for the polytomous response situation using (a) simultaneous logit models and (b) logit models for branching ratios. These results are illustrated in a detailed analysis of Census data on the educational attainment of white males. In addition, we have explored techniques for assessing the efficacy of longitudinal vs. cross-sectional studies using categorical data.

Many educational investigators are interested in providing a causal interpretation of the statistical relationships they find in the course of modeling a set of data. Goodman has recently proposed an analogue to path analysis for qualitative variables. We explored extensions of this work to nonrecursive systems of logit models, and have demonstrated that there is no way to separate reciprocal effects of variables on such systems.

Since many inferences from counted data are based upon statistics which have asymptotic chi-square distributions, it is important to determine the small sample properties associated with these statistics for a broad range of models under both null and non-null conditions. A related problem arises when investigators include or exclude variables or collapse categories based on the number of observations in the table. The usual asymptotic theory no longer applies and we have developed an asymptotic theory in this case for chi-square statistics based on estimated parameters.

Most large scale surveys dealing with educational problems are based on complex combinations and multiple levels of stratification and cluster sampling. The existing methodology for the analysis of categorical data using loglinear models is appropriate only for data arising out of simple random sampling. We have begun work to adapt this methodology for use in the analysis of complex sample survey data. Work on this project has focused primarily on the analysis of data from cluster samples, and has provided results on the adoption of the usual chi-square statistics to this form of sampling.

Many data sources in educational research involve both qualitative and quantitative variables. We have examined, in a preliminary

fashion, the adaptation of Dempster's general logit model for problems involving mixtures of qualitative and quantitative response structures. We have also considered the problem of model selection for loglinear models, and have conducted a moderate-sized Monte Carlo study to compare several model selection strategies proposed in the literature.

Other areas of research supported by this project have included graphical methods for the presentation of multivariate data, and the use of loglinear models for categorical paired-comparison situations, with ties and order effects.

This project has led directly to the preparation of 2 Ph.D. dissertations, Technical Reports, and 2 other manuscripts. Of these 4 have now been published, 7 others have been accepted or tentatively accepted for publication, and 2 are currently under review. A listing of publications and other manuscripts is appended to this report.

REFERENCES FOR INTRODUCTION

Bishop, Y.M.M., Fienberg, S.E. and Holland, P.W. [1975]. Discrete Multivariate Analysis: Theory and Practice. Cambridge, Massachusetts: The MIT Press.

Bock, R.D. [1975]. Multivariate Statistical Methods in Behavioral Research. New York: McGraw-Hill.

Cox, D.R. [1970]. Analysis of Binary Data. London: Methuen.

Fienberg, S.E. [1977]. The Analysis of Cross-classified Data, in preparation.

Gokhale, D.V. and Kullback, S. [1978]. The Information in Contingency Tables. New York: Marcel Decker.

Haberman, S.J. [1974]. The Analysis of Frequency Data. Chicago: University of Chicago Press.

Lindsey, J.K. [1973]. Inferences from Sociological Survey Data. Amsterdam: Elsevier.

McKinlay, S.M. [1975]. "The Design and Analysis of the Observational Study: A Review," J. Amer. Statist. Assoc., 70, 503-520.

Plackett, R.L. [1974]. The Analysis of Categorical Data. New York: Hafner Press.

II. DESCRIPTION OF RESEARCH

1. Age, Period, and Cohort Effects in the Analysis of Discrete Archival Data.

This research involved the collaborative efforts of Stephen Fienberg and William Mason of the Department of Sociology, University of Michigan, and has let to a lengthy publication entitled: "Identification and Estimation of Age-Period-Cohort Models in the Analysis of Discrete Archival Data" which appeared as a chapter in Sociological Methodology, 1979. The following description has been abstracted from the introduction to this chapter.

"This chapter discusses and resolves serveral issues arising out of a particular form of cohort analysis. The exposition is in terms of discrete data analysis, both because it may help those working with discrete data and because many of the general issues involved in the form of cohort analysis that concerns us can be elucidated with special clarity in the discrete data context."

"The version of cohort analysis we consider involves measurements on some dependent quantity conditional on the ages of the units of observation and the period at which the measurements were taken. We assume that replicated cross-sections are available and, in general, that repeated measurements on the same units of observation across replications are not available. The dates of data collection define the periods; and for each cross-section, measurements on age, on the dependent quantity, and possibly on other variables are available. Cohort membership is then defined by knowledge of age and the date or period at which that age was attained. The questions of

The questions of interest for the analyst are whether age, period, and cohort simultaneously determine the endogenous quantity -- and if so, how. We assume that whatever the explanation of the effect of cohort on the endogenous variable, no more direct measurement of the explanans than cohort membership itself is available for immediate purposes. We make similar assumptions about age and period. If more direct measures of the phenomena presumed to underlie the effects of age or period or cohort are available, then the problems that concern us do not arise."

"We have divided this chapter into six parts. (1) The Specification: This section presents the basic specification and sampling models we wish to consider for discrete response variables. (2) The 3 × 3 × 2 Case: We first consider the special case of three age groups, three periods, and a dichotomous response variable in order to develop the major points about identification, estimation, and goodness of fit as clearly as possible before proceeding to general cases. This section also illustrates for the 3 × 3 × 2 case how to arrange the data for computation with the two alternative algorithms (iterative proportional fitting and Newton-Raphson) discussed in this study. (3) The I × J × 2 Case: In this section we generalize the 3 × 3 × 2 case to the case of arbitrarily many age groups and periods. (4) The I × J × L Case: This section considers the case in which the response variable is polytomous. One interesting result of this section is that a polytomous response variable can sometimes be rearranged into a series of asymptotically independent dichotomies, so that the data can be treated as a series of I × J × 2 problems. (5) Multiple Age Groups and Uneven Spacing:

This section first considers the case of multiple age groups per period interval with age groups and periods evenly spaced. A new identification problem arises here. We indicate the solution to this problem and generalize the earlier results for estimation and degrees of freedom to this case. We then consider unevenly spaced periods. There currently is no satisfactory way to estimate age-period-cohort models for data with this feature. (6) Educational Attainment of White Males in America: We present an empirical analysis to illustrate both the theory set forth in preceding sections and the general nature of reasoning appropriate to the specification of age-period-cohort models. The data, which pertain to the educational attainment of white males, are from the 1940, 1950, 1960, and 1970 U.S. censuses".

2. Analysis and Design Problems Related to the Use of Longitudinal and Cross-sectional Qualitative Data.

This research was carried out by Stephen Fienberg and Richard R. Picard, a graduate student in the School of Statistics, who served as a Research Assistant on the project. The research work has resulted in a University of Minnesota - School of Statistics Technical Report (No. 321) entitled, "Designing surveys for measuring change in categorical data structures over time". This paper will be published in a volume entitled, Multivariate Analysis - V. This volume contains invited papers presented at the Fifth International Symposium on Multivariate Analysis, in June 1978. The following description of the research is abstracted directly from the paper.

"Statistical methods for the design and analysis of sample survey data have traditionally been geared toward the estimation of aggregate quantities for single points in time, such as population rates or averages, and the changes in these quantities from one point in time to another. The development of optimal or efficient sampling designs for such quantities, balancing costs of various sorts and accuracy of estimation, has often led to surveys which involve the repeared measurement of the same individuals or households over time".

"In this paper we consider a methodology for the analysis of categorical data structures over time based on loglinear models. In particular we consider situations where some sample individuals have complete longitudinal records, while others have less than complete records, due to absence from the sample. For concreteness we describe our approach in the relatively simple situation involving two dichotomous variables measured at two points in time, the so-called two-wave,

two-variable panel study. In this situation we can have data on individuals for both time points, for only the first, or for only the second. A sample design in this context invoives specifying which combination of the three forms of data is to be used and a mechanism for selecting the individuals for the three 'subsamples.'"

3. Systems of Logit Models and Path Diagrams for Causal Inference.

This research was carved out by Stephen S. Briar, who at the time was a graduate student in the School of Statistics, working as a Research Assistant under the direction of Stephen E. Fienberg. The research has led to a publication entitled "The Utility of Systems of Simultaneous Logistic Response Equations," which has appeared as a chapter in Sociological Methodology 1979. The following description is from the introduction to this paper.

"Econometricians have developed multiple-equation models to describe systems in which variables have simultaneous effects on each other. Sociologists have adapted these models to sociological problems. Until recently all these formulations have been for continuous variables -- that is, normal-theory models. Nerlove and Press develop what they refer to as multivariate log linear models for categorical response variables that may depend on continuous variables, and they present model specification and estimation techniques for a number of interesting situations. Goodman also presents multiple-equation logit models for the analysis of categorical data when there is a natural ordering of causal effects."

"Although analogies to the normal-theory simultaneous equation models have been drawn, I show here that there are important differences between the categorical case and the continuous-variable case. I show first that there is no identification problem in the discrete multiple-logit-equation case and second that there is no way to separate reciprocal effects of variables."

4. Small Sample Properties of Chi-square Statistics.

This research was carried out by Kinley Larntz, and has resulted in a publication entitled, "Small-sample comparisons of exact levels for Chi-squared Goodness-of-fit Statistics", Journal of the American Statistical Association 73 (1978). The abstract of this paper follows.

"The small-sample properties of three goodness-of-fit statistics for the analysis of categorical data are examined with respect to the adequacy of the asymptotic chi-squared approximation. The approximate tests based on the likelihood ratio and Freeman-Tukey statistics yield exact levels that are typically in excess of the nominal levels for moderate expected values. In contrast, the Pearson statistic attains exact levels that are quite close to the nominal values. The reason for the large number of rejections for the likelihood ratio and Freeman-Tukey statistics is related to their handling of small observed counts."

Larntz's results are also summarized in a review paper on the use of chi-squared statistics prepared as part of the project by Stephen Fienberg. This paper is entitled, "The Use of Chi-squared Statistics for Categorical Data Problems," and is scheduled to appear in the Journal of the Royal Statistical Society Series B in the first issue of 1979. The abstract reads as follows:

"The use of χ^2 statistics for categorical data problems was initiated by Karl Pearson, but it took several years before the asymptotic distribution of these statistics was well understood. The general structure of asymptotic results for χ^2 statistics is reviewed and the applicability of the general structure to a variety

of problems of practical interest is discussed. These problems include the use of χ^2 statistics in small-sample situations and in large sparse tables, in cluster sampling, and in cases where they do not have asymptotic χ^2 distributions."

An interesting feature of this review paper is the link it provides between this topic, and those described below in items 6 and 9.

5. Model For Mixtures of Qualitative and Quantitative Variables.

This aspect of the project was carried out by Michael Meyer, a graduate student in the School of Statistics, working as a Research Assistant under the joint direction of Stephen Fienberg and Kinley Larntz.

For most forms of statistical analysis we use models that distinguish between variables that are free to vary in response to controlled conditions, i.e. response variables, and variables that are regarded as fixed, either as in experimentation or because the context of the data suggests they play a determining or causal role in the situation under study, i.e. explanatory variables. This research explored the use of a recently proposed class of logit models for situations involving a mixture of continuous and discrete response variables.

Suppose $\underset{\sim}{Y}$ represents a set of multinomial response variables, $\underset{\sim}{X}$ a set of continuous response variables, and $\underset{\sim}{Z}$ a set of explanatory variables. Dempster [1977] proposed a general logit model for the joint density of $\underset{\sim}{Y}$ and $\underset{\sim}{X}$ given $\underset{\sim}{Z}$, $h(\underset{\sim}{Y}, \underset{\sim}{X}|\underset{\sim}{Z})$, i.e.

$$h(\underset{\sim}{Y}, \underset{\sim}{X}|\underset{\sim}{Z}) = f(\underset{\sim}{Y}|\underset{\sim}{Z})g(\underset{\sim}{X}|\underset{\sim}{Y}, \underset{\sim}{Z})$$

where $f(\underset{\sim}{Y}|\underset{\sim}{Z})$ is the density of $\underset{\sim}{Y}$ given $\underset{\sim}{Z}$, and $g(\underset{\sim}{X}|\underset{\sim}{Y}, \underset{\sim}{Z})$ is the density of $\underset{\sim}{X}$ given $\underset{\sim}{Y}$ and $\underset{\sim}{Z}$. The density $f(\underset{\sim}{Y}|\underset{\sim}{Z})$ is taken to be multinomial, where the dependency of $\underset{\sim}{Y}$ on $\underset{\sim}{Z}$ is given by the standard logistic response model, and $g(\underset{\sim}{X}|\underset{\sim}{Y}, \underset{\sim}{Z})$ is the multivariate normal density, i.e. $\underset{\sim}{X}$ given $\underset{\sim}{Y}$ and $\underset{\sim}{Z}$ has the $(Y\xi + Z\eta, \Sigma)$ distribution where $\underset{\sim\sim}{Y\xi} + \underset{\sim\sim}{Z\eta}$ is a set of linear combinations of the discrete response variables and the explanatory variables.

The computation problems associated with maximum likelihood estimates for this problem are complex. Thus attention was focused on a special case involving two response variables, one binomial and one normal (for example, suppose X is a measure of IQ, and Y indicates pass-fail on an examination). Meyer prepared a computer program to compute MLE's for this case, and used it for the analysis of several simulated data sets. The results are described in an internal memorandum, but have not been written up more formally.

At this point, we have concluded that more work needs to be done before we can comment on the utility of this approach for actual data.

6. Asymptotic Properties of Chi-squared Statistics in Large Sparse Contingency Tables.

This work involved the collaborative efforts of Kinley Larntz and Kenneth J. Koehler, who was a Research Assistant for the project and is now Assistant Professor of Statistics at Iowa State University. Their initial results were reported at the 1977 American Statistical Association meetings in Chicago. This initial paper has been revised and appears as School of Statistics Technical Report No. 327 "An Empirical Investigation of Goodness-of-Fit Statistics for Sparse Multinomials."

A summary of the paper reads:

"Traditional discussions of goodness-of-fit tests for multinomial data consider asymptotic chi-square properties under the assumption that all expected cell frequencies become large. However, this condition is not always satisfied and other asymptotic theories must be considered. For testing a specified simple hypothesis, Morris gave conditions for the asymptotic normality of the Pearson and likelihood ratio statistics when both the sample size and number of cells become large (even if the expected cell frequencies remain small). Monte Carlo techniques are used to examine the applicability of the normal approximations for moderate sample sizes with moderate numbers of cells."

This paper has been submitted to the Theory and Methods section of the Journal of the American Statistical Association.

Koehler submitted his doctoral dissertation in December 1977. It included the results mentioned above as well as a more theoretical treatment of the asymptotic properties of goodness-of-fit statistics

for large, sparse multinomials. Based on these theoretical aspects, Koehler has prepared a paper, "Distribution of the Likelihood Ratio Goodness-of-fit Statistic for Contingency Tables with Small Expected Frequencies." A summary of the paper reads:

"This paper considers the asymptotic normality of the likelihood ratio goodness-of-fit statistic for contingency table models under conditions which do not require the minimum expected frequency to become large. The traditional chi-squared theory is not applicable under these conditions. One difficulty in the application of the normal theory arises from the functional dependence of the asymptotic mean and variance on the unknown parameters defined by the true model. However, it is shown that under very general conditions a consistent estimate of the asymptotic variance is obtained by replacing the unknown parameters with these corresponding maximum likelihood estimates. This estimation method requires more stringent conditions to yield a consistent estimate of the asymptotic mean. Some results of a small Monte Carlo study are also presented."

This paper has been sumbitted to the Annals of Statistics.

In the Monte Carlo work examining the properties of the likelihood ratio statistic, a control variate technique was developed that made use of exact properties of the Pearson statistic. Details on the control variate technique were included in Koehler's dissertation. However, the exact moments of the Pearson statistic for alternative hypotheses was not previously known. In a paper, "A General Formula for Moments of the Pearson Goodness-of-fit Statistic for Alternatives," to be published in Biometrika, Koehler gives these results. A summary of that paper reads:

"A method for constructing formulae for the exact moments of the Pearson statistic from convolutions of polynomial functions of elementary multinomial probabilities is presented. A formula for the recursive generation of the polynomial coefficients is given and the coefficients required for the first six moments of the Pearson statistic are calculated. Formulae for the first four moments about zero are presented."

Larntz and Koehler have continued work on the practical application of the large sample results reported above. The emphasis is on obtaining results that will free the researcher from the onus of making decisions about collapsing over categories and/or variables in large, sparse contingency tables.

7. Selection Procedures for Hierarchical Loglinear Model Fitting.

The initial work in this area consisted of the preparation of several programs to automatically select an adequate model when fitting loglinear models to multidimensional contingency tables. The programs were written by Edward Bedrick, a Research Assistant working under the supervision of Kinely Larntz, and they perform a variety of stepwise procedures, as well as implementing Brown's method of partial and marginal association, and Wermuth's method of zero partial association.

Bedrick carried out a Monte Carlo experiment that used these programs on a variety of simulated data. Six separate probability structures were studied -- corresponding to situations ranging from complete independence and all pair wise association to varying degrees of saturation. Five structures involved five dimensional tables, while one was for a four dimensional table. In carrying out the Monte Carlo, complete records were kept on tape of all the steps followed in the selection procedures. The results of this Monte Carlo experiment have yet to be analyzed.

8. The Application of General Loglinear Model Results to New Problems.

The work was carried out by Stephen Fienberg, with some computational assistance from Michael Meyer. Two papers were prepared. The first gives a review of the general loglinear model results, with applications to multidimensional cross-classification, and the analysis of social interaction data. This manuscript, entitled "Using Loglinear Models to Analyze Cross-classified Categorical Data", has been tentatively accepted for publication in the Mathematical Scientist. A revised ms. was completed in December 1978, and it has been resubmitted.

The second paper applies the general loglinear results to a new problem involving paired comparisons problems. This paper, entitled "Loglinear Representation for Paired Comparison Models with Ties and Within-pair Order Effects" has been accepted for publication by Biometrics and will appear in the June 1979 issue. The introduction of this paper succinctly summarizes the value of the loglinear model approach:

"Davidson and Beaver have considered a generalization of the Bradley-Terry model for paired comparisons in the presence of ties and which examines the effect of the order of presentation of the objects within a pair. The present note shows how their model can be represented in loglinear form, and how the results for maximum likelihood estimation follow directly from known results for loglinear models fitted to categorical data. The advantages of the loglinear model approach to paired comparisons problems ... are three-fold:

(i) each variant on a standard model need not be considered de novo, since general results are available,

(ii) standard methods of solving likelihood equations can be used without new proofs of convergence,

(iii) further extensions to a class of models are often much easier to visualize when the models are represented in loglinear form."

9. Analysis of Qualitative Data from Complex Sample Surveys

This research was done by Stephen Brier, under the direction of Stephen Fienberg, and led directly to Brier's (1979) doctoral dissertation in the School of Statistics entitled, "Categorical Data Models for Complex Data Structures."

A question often asked in practice is: How inappropriate is the use of the usual goodness-of-fit statistics and the standard χ^2 reference distributions in situations where the data come from a cluster sample rather than the simple random sampling model implied by the multinomial distribution? Brier answers this question by exploring a special model for cluster structures. His key result is that in large samples a multiple of the usual chi-squared statistic can be used to test for goodness-of-fit by referring it to the standard χ^2 reference distribution. The "multiple" is a function of the data involving a measure of within cluster homogeneity. Several generalizations of this result have already been considered.

Brier plans to prepare several manuscripts for publication from the research that went into this dissertation. A preliminary version of one of these is available under the title, "Discrete data models with random effects."

10. Graphical Methods for the Presentation and Analysis of Multivariate Data.

This part of the project has involved reviews of the statistical literature of the graphical representation of data, and of the use of graphical methods more broadly. The focus has not been on categorical data. Two formal products are in hand. The first of these grew out of a General Methodology Lecture presented by Stephen Fienberg at the Annual Meetings of the American Statistical Association in August, 1977. A written version based on part of that lecture was available as School of Statistics Technical Report No. 304, under the title of "Graphical Methods in Statistics." The abstract of this paper, which is currently under review in revised form by the American Statistician, reads as follows:

"Graphical methods have played a central role in the development of statistical theory and practice. This presentation briefly reviews some of the highlights in the historical development of statistical graphics, and gives a simple taxonomy which can be used to characterize the current use of graphical methods. This taxonomy is used to describe the evolution of the use of graphics in some major statistical and related scientific journals.

Some recent advances in the use of graphical methods for statistical analysis are reviewed, and several graphical methods for the statistical presentation of data are illustrated, including the use of multi-colored maps.

The second paper was prepared by Alan Izenman, who is currently a visiting faculty member in the Department of Applied Statistics at the University of Minnesota, and is available as School of

Statistics Technical Report No. 335, "Developments in Statistical Graphics 1960-1980." The abstract reads as follows:

"This paper presents an historical perspective on the development of graphical methods in data analysis during the twenty years from 1960 to 1980. Three main periods of development are identified: pre-1960; the 1960's; and the 1970's. Discussion of various types of univariate and multivariate data graphics is given including methods for probability plotting. Possible future directions for statistical graphics are also mentioned."

This paper will appear in the Proceedings of the First General Conference on Social Graphics.

Together these two reviews suggest an agenda for future research in this area.

APPENDIX: RESEARCH PAPERS PREPARED

A. Published Papers.

1. Stephen S. Brier [1978]. "The Utility of Systems of Simultaneous Logistic Response Equations." Sociological Methodology 1979, 119-129.

2. Stephen E. Fienberg [1978]. "A Note on Fitting and Interpreting Parameters in Models for Categorical Data." Sociological Methodology 1979, 112-118.

3. _____ and William M. Mason [1978]. "Identification and Estimation of Age-Period-Cohort Models in the Analysis of Discrete Archival Data." Sociological Methodology 1979, 1-67.

4. Kinley Larntz [1978]. "Small-Sample Comparison of Exact Levels of Chi-Squared Goodness-of-Fit Statistics." Journal of the American Statistical Association, 73, 253-263.

B. Papers Accepted or Tentatively Accepted for Publication.

5. Stephen E. Fienberg [1979]. "The Use of Chi-Squared Statistics for Categorical Data Problems." Journal of the Royal Statistical Society, Series B, 41, (in press).

6. Stephen E. Fienberg [1979]. "Log Linear Representation for Paired Comparison Models with Ties and Within-Pair Order Effects." Biometrics, 35, (in press).

7. _____ "Using Loglinear Models to Analyse Cross-Classified Categorical Data." Mathematical Scientist, (tentatively accepted).

8. _____ "Graphical Methods in Statistics." American Statistician, (tentatively accepted).

9. _____ and Richard R. Picard [1979]. "Designing Surveys for Measuring Change in Categorical Data Structures Over Time." To appear in Multivariate Analysis - V. (edited by P.R. Krishnaiah).

10. Alan J. Izenman [1979]. "Developments in Statistical Graphics 1960-1980." To appear in Proceedings of First General Conference on Social Graphics.

11. Kenneth J. Koehler [1979]. "A General Formula for Moments of the Pearson Goodness-of-Fit Statistic for Alternatives."

C. Doctoral Dissertations at University of Minnesota.

12. Stephen S. Brier [1979]. Categorical Data Models for Complex Data Structures.

13. Kenneth J. Koehler [1978]. Goodness-of-Fit Statistics for Large Sparse Multinomials.

D. Papers Submitted for Publication.

14. Kenneth J. Koehler. "Distribution of the likelihood ratio goodness-of-fit statistic for contingency tables with small expected frequencies." Submitted to the Annals of Statistics.

15. _____ and Kinley Larntz. "An Empirical Investigation of Goodness-of-Fit Statistics for Sparse Multinomials." Submitted to the Journal of the American Statistical Association.

E. Other Manuscripts.

16. Stephen S. Brier. "Discrete Data Models with Random Effects." (March 16, 1978).

17. Kenneth J. Koehler and Kinley Larntz. "Goodness-of-fit Statistics for Large Sparse Multinomials." (August, 1977).